

# T3DB: a comprehensively annotated database of common toxins and their targets

Emilia Lim<sup>1</sup>, Allison Pon<sup>1</sup>, Yannick Djoumbou<sup>1</sup>, Craig Knox<sup>1</sup>, Savita Shrivastava<sup>1</sup>, An Chi Guo<sup>1</sup>, Vanessa Neveu<sup>1</sup> and David S. Wishart<sup>1,2,3,\*</sup>

<sup>1</sup>Department of Computing Science, <sup>2</sup>Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada T6G 2E8 and <sup>3</sup>National Institute for Nanotechnology, 11421 Saskatchewan Drive, Edmonton, AB, Canada T6G 2M9

Received August 15, 2009; Accepted October 11, 2009

## ABSTRACT

In an effort to capture meaningful biological, chemical and mechanistic information about clinically relevant, commonly encountered or important toxins, we have developed the Toxin and Toxin-Target Database (T3DB). The T3DB is a unique bioinformatics resource that compiles comprehensive information about common or ubiquitous toxins and their toxin-targets into a single electronic repository. The database currently contains over 2900 small molecule and peptide toxins, 1300 toxin-targets and more than 33 000 toxin-target associations. Each T3DB record (ToxCard) contains over 80 data fields providing detailed information on chemical properties and descriptors, toxicity values, protein and gene sequences (for both targets and toxins), molecular and cellular interaction data, toxicological data, mechanistic information and references. This information has been manually extracted and manually verified from numerous sources, including other electronic databases, government documents, textbooks and scientific journals. A key focus of the T3DB is on providing ‘depth’ over ‘breadth’ with detailed descriptions, mechanisms of action, and information on toxins and toxin-targets. T3DB is fully searchable and supports extensive text, sequence, chemical structure and relational query searches, similar to those found in the Human Metabolome Database (HMDB) and DrugBank. Potential applications of the T3DB include clinical metabolomics, toxin target prediction, toxicity prediction and toxicology education. The T3DB is available online at <http://www.t3db.org>.

## INTRODUCTION

A toxic substance is a small molecule, peptide, or protein that is capable of causing injury or death. Toxins, both natural and man-made, represent an important class of poisonous compounds that are ubiquitous in nature, in homes and in the workplace. Common toxins include pollutants, pesticides, preservatives, drugs, venoms, food toxins, cosmetic toxins, dyes and cleaning compounds. Indeed, common toxins are so ubiquitous that most adults will have been exposed to or will have ‘bioaccumulated’ several hundred of these compounds during their lifetime. While the concentration of most toxins in adults is generally below hazardous levels, their ubiquity essentially makes them an integral part of the human metabolome (1). This fact, combined with the recent advances in chemical detection technologies for monitoring low abundance compounds, has greatly increased the need to identify, quantify, classify and understand these common toxins (2).

Though a great deal of information has been acquired about toxic substances and their mechanism of action over the past 100 years, much of this information tends to be highly dispersed and of limited accessibility—often being confined to specialized textbooks, obscure government documents or subscription-requiring scientific journals. Fortunately, this situation is beginning to change. Over the past two years a number of excellent, web-accessible toxic substance databases have started to appear, such as the Animal Toxin Database (ATDB) (3), SuperToxic (4), Aggregated Computational Toxicology Resource (ACToR) (5) and the Comparative Toxicogenomics Database (CTD) (6). In particular, ACToR and SuperToxic provide bioassay data and chemical structure information for a very large number of industrial or pharmaceutically interesting chemicals (>60 000 for SuperToxic, >500 000 for ACToR). ATDB with 3844 peptide toxins and CTD with 5056 chemicals are

\*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: [david.wishart@ualberta.ca](mailto:david.wishart@ualberta.ca)

somewhat more specialized, offering data on channel targets (ATDB) or chemical-gene interactions (CTD) for a much smaller number of better-understood toxins.

Each of these databases addresses the needs of certain communities such as animal physiologists (ATDB), toxicogenomics specialists (CTD), environmental or industrial regulators (ACToR) or medicinal chemists interested in toxicity prediction (SuperToxic). However, compared to a number of other widely used biochemical or metabolomic databases such as Human Metabolome Database (HMDB) (1), DrugBank (7) or PharmGKB (8) most of today's online toxin or toxic compound databases are relatively lightly annotated, with fewer than 10 data fields per compound. Furthermore, the data model adopted by many databases seems to be focused on providing a little information about everything (including somewhat exotic or extremely rare compounds) at the expense of including a lot of information about those toxins that nearly everyone encounters. Consequently it is difficult to find answers to common questions such as: How does this poison work? What enzymes, receptors or organs does this toxin target? Where is this chemical found? How does this toxin get into the body? What are the toxic or lethal levels for this compound? How can I treat someone who has been exposed to this substance? Where can I find more information about this chemical? In other words, the information contained in these resources does not fully address the needs of physicians, toxicologists, clinical chemists, metabolomics researchers, educators and the general public. Indeed, members of these communities typically need in-depth, biological, chemical and mechanistic information about clinically relevant or commonly encountered toxins. In an effort to address these needs we have developed the Toxin and Toxin-Target Database (T3DB).

## DATABASE DESCRIPTION

T3DB is a multi-purpose, freely available, online toxicology resource that provides in-depth, quantitative and molecular-scale information about toxins and their associated targets. In terms of its layout, content and depth of coverage, the T3DB is modeled closely after two other cheminformatic or metabolomics resources previously developed in our laboratory: the HMDB and DrugBank. As the name indicates, the T3DB is primarily intended to be a database that links toxins with their biological (usually protein) targets. This molecular interaction information is further supplemented with detailed descriptions of the toxin's mechanism of action, its metabolism in the human body, its lethal or toxic dose levels, its potential carcinogenicity, exposure sources, symptoms or health effects and suggested treatment options. In total each toxin listed in T3DB contains more than 80 data fields covering the structure, chemistry, physico-chemical properties, toxicology, route of delivery, treatment and target protein/macromolecule properties. These data fields are listed in Table 1. T3DB currently contains over 2900 toxin entries corresponding to more than 34 000 different synonyms. These toxins are further connected to

**Table 1.** Summary of the data fields or data types found in each ToxCard

Toxin and medical information	Toxin target—protein/enzyme information
Common name	Name
Compound type	Mechanism of action (in humans)
Description	Description
Synonyms/IUPAC name	Synonyms
Chemical formula	Gene name
Chemical structure	Protein sequence
CAS registry number	Number of residues
InChi identifier	Molecular weight
PubChem compound ID	Theoretical pI
KEGG compound ID	GO classification
UniProt ID	General function
OMIM ID	Pathways
ChEBI ID	Reactions
BioCyc ID	Signals
SuperToxic ID	Transmembrane regions
CTD ID	Essentiality
STITCH ID	Domain function
DrugBank ID	GenBank ID protein
PDB ID	UniProtKB ID
ACToR ID	Cellular location
Wikipedia link	Gene sequence
Monoisotopic mass	GenBank gene ID
MOL file	GeneCard ID
PDB file	GenAtlas ID
SMILES	HGNC ID
Appearance	Chromosome location
Melting point	Locus
Boiling point	SNPs
Density	General references
Solubility	
Predicted LogP	
Specific gravity	
Flash point	
Vapour pressure	
Route of exposure	
Mechanism of action	
Metabolism	
Toxicity values (experimental LD50s and LC50s)	
Lethal dose (for humans)	
Carcinogenicity (IARC classification)	
Uses/sources	
Minimum risk level	
Health effects	
Symptoms	
Treatment	
General references	
Protein targets	

some 1300 protein targets through almost 33 500 toxin and toxin-target bonds which, in turn, are supported by more than 3100 references. The entire database, including text, sequence, structure and image data, occupies nearly 16 Gigabytes of data—most of which can be freely downloaded.

The diversity of data types, required breadth of knowledge and widespread nature of the required data made the assembly of the T3DB both difficult and time-consuming. The T3DB represents the accumulation of both qualitative and quantitative information derived from a variety of sources, including scientific journals, government records, chemical safety documents

(i.e. MSDS or Material Safety Data Sheets), textbooks, encyclopedias, online protein and protein structure databases, on-line metabolomic databases, electronic drug and pharmaceutical databases as well as existing toxin and toxicology databases. To compile, confirm and validate this comprehensive collection of data, more than a dozen toxicology, pharmacology and biochemistry textbooks, several thousands journal articles, nearly 20 different electronic databases, and at least 20 in-house or web-based programs were individually searched, accessed, compared, written, or run over the course of the past year. The team of T3DB contributors and annotators consisted of eight bioinformaticians with dual training in computing science and molecular biology/chemistry.

The T3DB is designed to be a fully searchable web resource with many built-in tools and features for viewing, sorting and extracting toxin and toxin-target annotation, including structures and gene and protein sequences. A screenshot montage illustrating the types of viewing and searching options available is shown in Figure 1. Detailed instructions on where to locate and how to use these browsing/search tools are provided on the T3DB homepage. As with any web-enabled database, the T3DB supports standard text queries through the text search box located on the home page. It also offers general database browsing using the "Browse" button located in the T3DB navigation bar. To facilitate browsing, the T3DB is divided into synoptic summary tables which, in turn, are linked to more detailed 'ToxCards'—in analogy to the very successful 'DrugCard' concept found in DrugBank (7). All of the T3DB's summary tables can be rapidly browsed, sorted or reformatted in a manner similar to the way PubMed abstracts may be viewed. Clicking on the ToxCards button found in the leftmost column of any given T3DB summary table opens a webpage describing the toxin of interest in much greater detail. Each ToxCards entry contains over 80 data fields, with ~50 data fields devoted to chemical and toxicological/medical data and ~30 data fields (each) devoted to describing the toxin target(s).

A ToxCards begins with various identifiers and descriptors (names, synonyms, compound description, structure image, related database links and ID numbers), followed by additional structure and physico-chemical property information. The remainder of data on the toxin is devoted to providing detailed toxicity and toxicological data, including route of delivery, mechanism of action, medical information and toxicity measurements. All of a toxin's targets are also listed within the ToxCards. Each of these targets are described by some 30 data fields that include both chemical and biological information, as well as details on their role in the mechanism of action of the toxin. In addition to providing comprehensive numeric, sequence and textual data, each ToxCards also contains hyperlinks to other databases, abstracts, digital images and interactive applets for viewing molecular structures.

A key feature that distinguishes the T3DB from other on-line toxin or toxicology resources is its extensive support for higher-level database searching and selecting functions. In addition to the data viewing and sorting

features already described, the T3DB also offers a local BLAST (9) search that supports both single and multiple sequence queries, a boolean text search based on KinoSearch (<http://www.rectangular.com/kinosearch/>), a chemical structure search utility based on ChemAxon's MarvinView, and a relational data extraction tool similar to that found in DrugBank and the HMDB (1). These can all be accessed via the database navigation bar located at the top of every T3DB page.

T3DB's simple text search box (located at the top of most T3DB pages) supports text matching, text match rankings, mis-spellings (offering suggestions for incorrectly spelled words) and highlights text where the word is found. In addition to this simple text search, T3DB's TextQuery function uses the same KinoSearch engine, but also supports more sophisticated text querying functions (Boolean logic, multi-word matching and parenthetical groupings) as well as data-field-specific queries such as finding the query word only in the 'Compound Source' field. Additional details and examples are provided on the HMDB's TextQuery page.

T3DB's sequence searching utility (SeqSearch) allows users to search through T3DB's collection of 1300 known (human) toxin targets. This service potentially allows users to identify both orthologous and paralogous targets for known toxins or toxin targets. It also facilitates the identification of potential targets from other animal species. With SeqSearch, gene or protein sequences may be searched against the T3DB's sequence database of identified toxin-target sequences by pasting the FASTA formatted sequence (or sequences) into the SeqSearch query box and pressing the 'submit' button.

T3DB's structure similarity search tool (ChemQuery) can be used in a similar manner as its SeqSearch tool. Users may sketch (through ChemAxon's freely available chemical sketching applet) or paste a SMILES string (10) of a query compound into the ChemQuery window. Submitting the query launches a structure similarity search that looks for common substructures from the query compound that matches the T3DB's database of known toxic compounds. Users can also select the type of search (exact or Tanimoto score) to be performed. High-scoring hits are presented in a tabular format with hyperlinks to the corresponding ToxCards (which in turn links to the targets). The ChemQuery tool allows users to quickly determine whether their compound of interest is a known toxin or chemically related to a known toxin and which target(s) it may act upon. In addition to these structure similarity searches, the ChemQuery utility also supports compound searches on the basis of chemical formula and molecular weight ranges.

The T3DB's data extraction utility (Data Extractor) employs a simple relational database system that allows users to select one or more data fields and to search for ranges, occurrences or partial occurrences of words, strings, or numbers. The data extractor uses clickable web forms so that users may intuitively construct SQL-like queries. Using a few mouse clicks, it is relatively simple to construct complex queries ('find all toxins that target acetylcholinesterase and are pesticides') or to build



The figure is a screenshot montage of the Toxin, Toxin-Target Database (T3DB) interface. It includes the following components:

- Top Header:** The T3DB logo and navigation links: Home, Browse, Search, About, Downloads, Contact Us.
- Search Bar:** A search input field with the placeholder text "Search: Search T3DB" and a "Search" button.
- Browsing toxins:** A section showing a list of toxins with columns for T3DB ID, Name, CAS Number, Formula, Weight, Structure, Compound Type, and Health Effects. The list includes Hexachlorobutadiene, Aldrin, DDT, and Toxaphene.
- Showing toxin card for Toxaphene (T3D0031):** A detailed view of a specific toxin card, including its version, creation date, update date, accession number, name, compound type, description, synonyms, chemical IUPAC name, chemical formula, chemical structure, CAS registry number, InChI identifier, PubChem compound ID, and KEGG ID.
- Data Extractor:** A tool for extracting data from the database, featuring a legend for toxin and target fields, and a list of fields to be extracted (e.g., CAS Number, Name, Compound Type, Solubility, Molecular Weight, Theoretical pI).
- ChemQuery:** A window for chemical structure similarity search, showing a search type (Tanimoto Similarity), similarity threshold (0.7), molecular weight filter, and maximum results returned (100).

**Figure 1.** A screenshot montage of the Toxin, Toxin-Target Database (T3DB) showing several of T3DB's search and display tools describing the toxin Toxaphene. Not all fields are shown. Clockwise from top left: Toxin Browse view; ToxCards for Toxaphene; chemical structure similarity search for Toxaphene; T3DB ChemQuery window; T3DB data extractor for toxin and toxin-target data fields.

a series of highly customized tables. The output from these queries is provided in HTML format with hyperlinks to all associated ToxCards.

## DATABASE IMPLEMENTATION

T3DB has a perl-based web-friendly front-end attached to a sophisticated MySQL relational database (version: 5.0.51a). The database is maintained on an Apple

XServe located on the University of Alberta campus. All data is entered directly onto the database's Ruby On Rails web-based Laboratory Information Management System (T3-LIMS). Each ToxCards and TargetCard has an edit page, which allows administrators to manually make changes to the database entries. The public user interface and the LIMS both read from the same database, and there is no intermediate database for rendered HTML pages.

All structures in the T3DB are stored in a centralized structure hub. This hub is a RESTful web resource that automatically stores and updates chemical properties such as molecular weight, solubility and log P. Additionally, the hub renders the structure images and thumbnails visible on the public site. T3DB's structure hub uses a combination of JChem Base (from ChemAxon) and Ruby on Rails running in a jRuby environment on a GlassFish server. The centralized nature of this hub helps to maintain consistency for all structures stored in T3DB. Whenever a structure is changed or fixed, all properties are automatically re-calculated and made available on the public site.

### CRITERIA FOR INCLUSION IN T3DB

Unlike the situation for drugs where formal approval processes and tight regulations allows one to easily identify whether a compound is a drug or not, the situation for toxins is not so simple. First, there is the issue of defining exactly what a toxin is. Normally, the word 'toxin' refers to poisonous substances produced by living organisms, while the term 'poison' refers to toxic substances that are both naturally occurring and man-made. Given that many man-made toxic substances are also produced by living organisms or are the byproducts of living organisms (albeit at low concentrations), this distinction seems a little arbitrary. Therefore we have chosen to use the term toxin to refer to both man-made and naturally produced substances that exhibit some form of toxicity (i.e. an acute reaction, injury or death). In constructing the T3DB a distinction between 'all possible' toxins and 'common' toxins also had to be made. Technically almost all of the 38-million known chemical substances (11) could be toxic if consumed, injected or inhaled in sufficient quantities. Yet only a tiny fraction of known chemicals ever make it out of the laboratory and into everyday use (perhaps <20 000 compounds). Therefore 'common toxins' would have to be defined as those compounds that can be found in the home, the environment or the workplace and which of have had a recorded medical consequence (i.e. acute reaction, injury or death).

While this definition may seem reasonable, some further constraints also need to be applied to make a 'common toxin' database useful. For instance, essential compounds such as vitamins, sugar, oxygen and even water can be prove to be toxic if consumed/inhaled in extreme excess. Yet their inclusion in a toxin database like T3DB would be certainly stretch the limits of believability. Therefore we chose to further refine the inclusion criteria to limit ourselves to those compounds that have been routinely identified as hazardous in relatively low concentrations (<1 mM for some, <1  $\mu$ M for others) and which appear on multiple toxin/poison lists provided by government agencies such as TOXNET (12) or the toxicological and medical literature. In each case, the toxicity of each compound was assessed by examining the available toxicity measurements and health effects, such as minimum lethal dose, LD50, LC50 values and carcinogenicity.

Furthermore, because many common toxic substances, such as cleaning agents, gasoline, kerosene, glue, paint

thinners are complex mixtures of many compounds, we also limited T3DB to include only those substances where specific toxic components could be identified and where the chemical structures to those components were known. As a result, common or name-brand household cleaning products are not included in T3DB but their toxic components are. Obviously these criteria may lead to the exclusion of some rare or interesting toxic compounds or toxic mixtures, but our focus for T3DB was on creating a chemically-oriented database that could be broadly useful, comprehensively annotated and easily managed.

### QUALITY ASSURANCE, COMPLETENESS AND CURATION

As has been previously noted, T3DB toxins were identified using a number of methods, including literature surveys, data mining of other databases, using toxicology textbooks and examining lists of controlled or banned substances. The toxic compounds found using these sources were also used to derive additional substances that were toxic by relation (such as metal salts or structurally similar compounds) and which can be found in the household, workplace or environment. Once identified the toxins were further assessed using the 'common toxin' inclusion criteria described above. In order to ensure both completeness and correctness, each toxin record entered into T3DB was reviewed and validated by a member of the curation team after being annotated by another member. Additional spot checks were routinely performed on each entry by other members of the curation group. Several software packages including text-mining tools, chemical parameter calculators and protein annotation tools were developed, modified and used to aid in data entry and data validation. This includes search tools such as BioSpider (13) and PolySearch (14), which collate and display text (as well as images) from multiple sources allowing T3DB's curators to compare, assess, enter and correct toxin and toxin-target information.

Much of the annotation in T3DB was done manually, especially in areas regarding toxicology and biology, such as mechanisms of action, route of delivery, health effects and target identification. This qualitative and quantitative information was derived from a variety of sources, including scientific journals, government records, Material Safety Data Sheets (MSDS), textbooks, encyclopedias and existing toxicology databases. For example, we frequently used existing chemical-protein or chemical gene interaction databases such as the CTD and STITCH (15) to identify potential toxin targets. Once these potential targets were identified, the corresponding references were manually reviewed by multiple curators to confirm which of the genes/proteins were true targets. In annotating the toxic action of drugs, we often used the drug mechanism information found in T3DB's sister database DrugBank. To facilitate and monitor the data entry process, all of T3DB's data is entered into a centralized laboratory information management system

(LIMS), allowing all changes and edits to the T3DB to be monitored, time-stamped and automatically transferred.

## LIMITATIONS AND CONCLUSION

Relative to other recently published toxic substance databases, T3DB probably has the smallest number of toxins or poisons in its collection (~2900 in T3DB versus ~500 000 compounds in ACToR). This situation was quite deliberate as T3DB was designed to be a database for common toxins as opposed to a database for all known toxic substances. While T3DB's breadth of coverage is limited, its depth of coverage is not. With more than 80 data fields for each toxin and with more than 16 Gb of data, we believe T3DB offers a considerable amount of useful chemical, biological and toxicological data. As with most toxic substance databases T3DB is still a work in progress. For instance, some data fields for some toxins are still 'empty'. This usually indicates that the data has yet to be measured or the information has yet to be reported in the literature. In some cases, the data probably exists, but our annotation team has yet to locate or validate it. In addition to the existence of empty data fields, T3DB is also missing some compounds. Currently more than 500 compounds are still on the T3DB 'to do' list and will be added (at a rate of ~50 compounds a week) to the database over the coming months. Furthermore, new toxins are always being identified and these will need to be added as they are reported (if sufficient information exists). Because T3DB strives for in-depth coverage, not all toxic compounds make it to the database. Those that lack any significant biological or toxicological data (i.e. the claim for toxicity is not supported by any study or biomedical report) or those that do not have a known chemical structure will not be included in the database.

Despite these caveats and limitations, we believe T3DB represents a new and important model for toxic substance databases. With a focus on depth as opposed to breadth, we are hopeful that the rich content in T3DB will allow researchers, educators and members of the general public to uncover answers to many common questions about toxins and poisons. Likewise, with its unique emphasis on 'common' substances (i.e. those that can normally be detected with modern analytical methods) we believe that the T3DB should also prove to be a valuable resource in toxico-metabolomics and clinical toxicology research. The fact that T3DB provides a wide range of chemical and biological data as well as a broad collection of search and display tools also represents a significant new development for toxic substance databases. We are certainly hoping that some of these ideas may find their way into other toxic substance databases. Over the coming two years we also plan to bring in additional data (spectroscopic, tissue/biofluid concentration data) and to add new features (spectral matching, pathway images, target prediction) to this database, with the intent of making T3DB a comprehensive, one-stop-shop for toxicological and toxico-metabolomics research.

## ACKNOWLEDGEMENT

The authors wish to thank Igor Sinelnikov, Edison Dong and Paul Huang for the help in validating T3DB's structure data.

## FUNDING

Alberta Advanced Education and Technology (AAET); The Canadian Institutes of Health Research (CIHR); Genome Alberta, a division of Genome Canada. Funding for open access charge: Canadian Institutes of Health Research.

*Conflict of interest statement.* None declared.

## REFERENCES

- Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.
- Mattingly, C.J. (2009) Chemical databases for environmental health and clinical research. *Toxicol. Lett.*, **186**, 62–65.
- He, Q.Y., He, Q.Z., Deng, X.C., Yao, L., Meng, E., Liu, Z.H. and Liang, S.P. (2008) ATDB: a uni-database platform for animal toxins. *Nucleic Acids Res.*, **36**, D293–D297.
- Schmidt, U., Struck, S., Gruening, B., Hossbach, J., Jaeger, I.S., Parol, R., Lindequist, U., Teuscher, E. and Preissner, R. (2009) SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.*, **37**, D295–D299.
- Judson, R., Richard, A., Dix, D., Houck, K., Elloumi, F., Martin, M., Cathey, T., Transue, T.R., Spencer, R. and Wolf, M. (2008) ACToR—aggregated computational toxicology resource. *Toxicol. Appl. Pharmacol.*, **233**, 7–13.
- Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C. and Mattingly, C.J. (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.*, **37**, D786–D792.
- Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B. and Hassanali, M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
- Sangkul, K., Berlin, D.S., Altman, R.B. and Klein, T.E. (2008) PharmGKB: understanding the effects of individual genetic variants. *Drug Metab. Rev.*, **40**, 539–551.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Weininger, D. (1988) SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–38.
- Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J. and Bryant, S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
- Wexler, P. (2004) The US National Library of Medicine's Toxicology and Environmental Health Information Program. *Toxicology*, **198**, 161–168.
- Knox, C., Shrivastava, S., Stothard, P., Eisner, R. and Wishart, D.S. (2007) BioSpider: a web server for automating metabolome annotations. *Pac. Symp. Biocomput.*, 145–156.
- Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S. and Wishart, D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.
- Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. and Bork, P. (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.